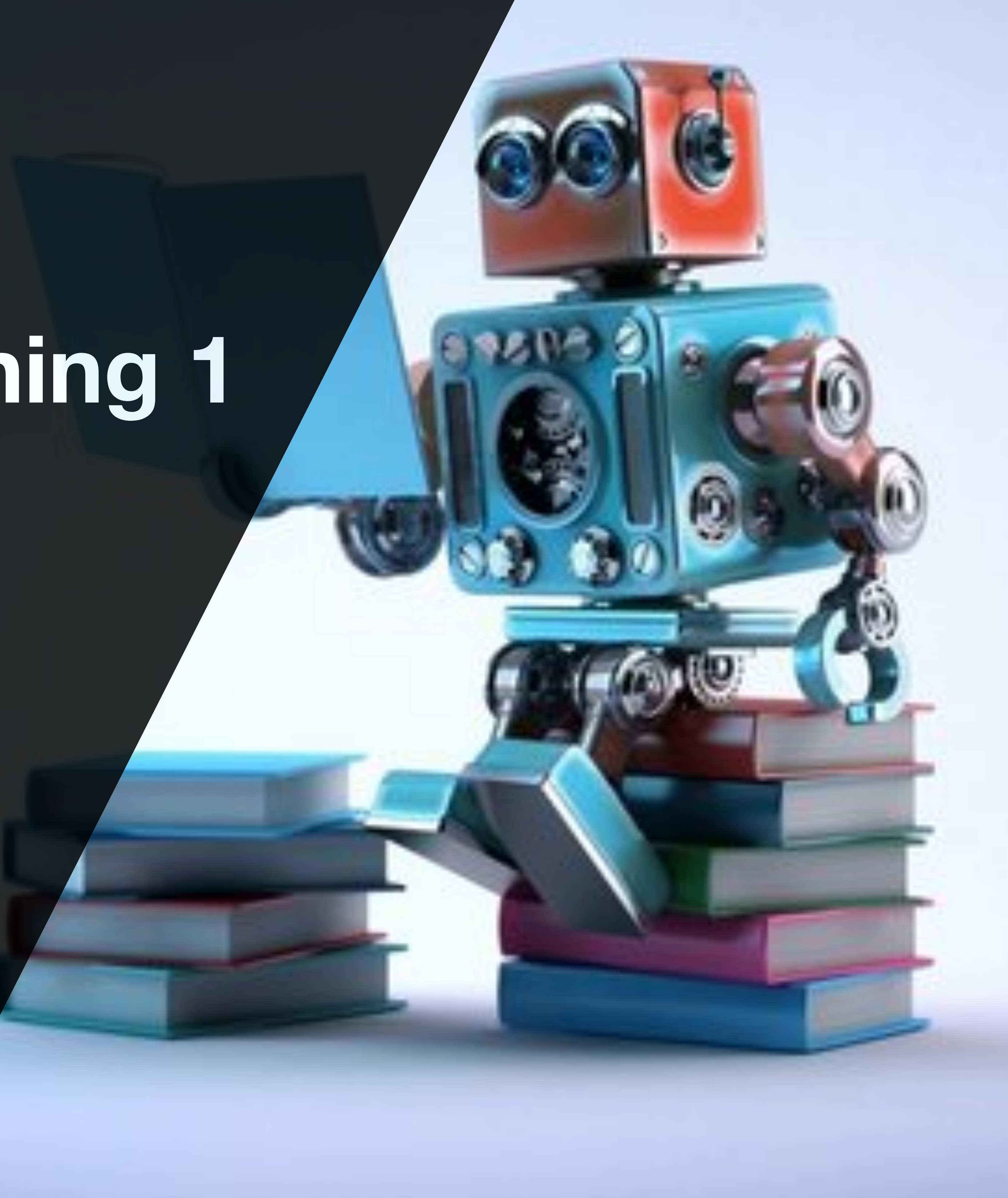
Machine Learning 1

Lecture 11.2 - Kernel Methods
Kernel Trick - Valid Kernels

Erik Bekkers

(Bishop 6.2)



Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Kernel Trick/Kernel substitution

- Formulate your optimization problem in such a way that the input vectors $\mathbf{x}_n$ enter only in the form of scalar products:

$$\mathbf{x}_n^T \mathbf{x}_n \qquad \text{(or when using basis functions } \boldsymbol{\phi}(\mathbf{x}_n)^T \boldsymbol{\phi}(\mathbf{x}_n)\text{)}$$

- Replace all instances of $\mathbf{x}_n^T \mathbf{x}_m$ with a kernel function

$$k(\mathbf{x}_n, \mathbf{x}_m) = K_{nm}$$

→ Store in Gram matrix $N \times N$

- Kernel $k(\mathbf{x}_n, \mathbf{x}_m)$ corresponds to a scalar product in some (possibly infinite dimensional) feature space.

- Valid kernel: Gram Matrix $\mathbf{K}$ must be symmetric positive semi definite for all possible choices of $\{\mathbf{x}_n\}_{n=1}^N$

$\underline{z}^T K \underline{z} \geq 0, \ \underline{z} \in \mathbb{R}^N, \ K \in \mathbb{R}^{N \times N}$

$k(\underline{x}, \underline{x}') = \phi(x)^T \phi(x')$: $\quad \underline{z}^T K \underline{z} = \underline{z}^T \Phi \Phi^T \underline{z} = (\Phi^T \underline{z})^T (\Phi^T \underline{z}) = \|\Phi^T \underline{z}\|^2 \geq 0$

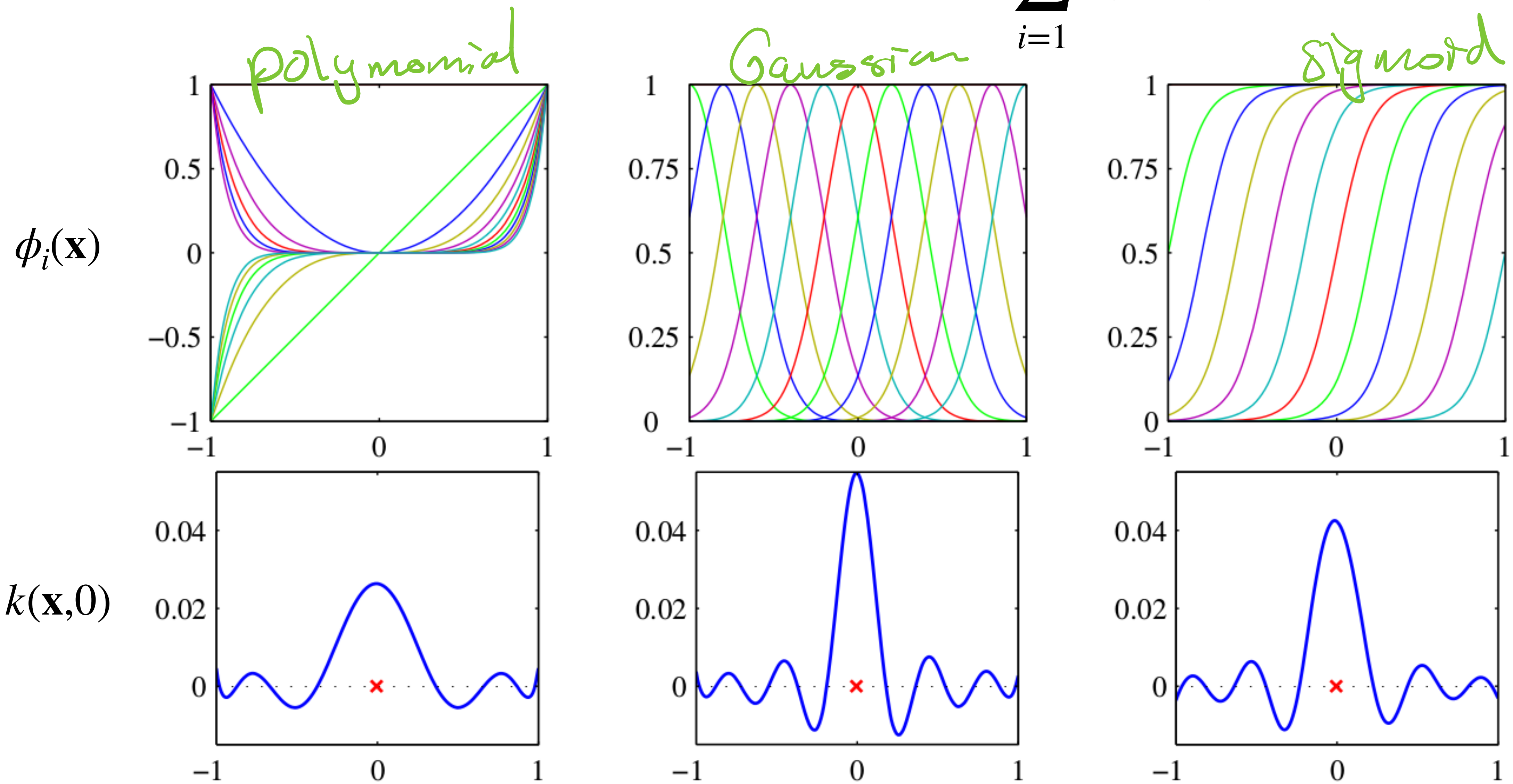
Constructing valid kernels

- Construct kernel from explicit set of basis functions:

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}') = \sum_{i=1}^{M} \phi_i(\mathbf{x})\phi_i(\mathbf{x}')$$

$$\left(\begin{aligned} \boldsymbol{\Sigma} &= \mathbf{U}^T \boldsymbol{\Lambda} \mathbf{U} \\ \boldsymbol{\Sigma}^{1/2} &= \mathbf{U}^T \boldsymbol{\Lambda}^{1/2} \mathbf{U} \\ \boldsymbol{\psi}(\mathbf{x}) &= \boldsymbol{\Sigma}^{1/2} \boldsymbol{\phi}(\mathbf{x}) \end{aligned} \right)$$

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\psi}(\mathbf{x}') = \sum_{i=1}^{M} \psi_i(\mathbf{x})\psi_i(\mathbf{x}')$$



Deriving the corresponding feature vector

- For every positive definite kernel there exists $\boldsymbol{\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^M$ such that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x})$$

- Depending on the kernel, M can be infinite!

- In general difficult to retrieve the corresponding $\boldsymbol{\phi}(\mathbf{x})$ for a given kernel

Example: polynomial kernel

- Polynomial kernel of order $M = 2$ for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$

$$k(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T\mathbf{z})^2 = (1 + x_1 z_1 + x_2 z_2)^2 = (1 + x_1 z_1 + x_2 z_2)(1 + x_1 z_1 + x_2 z_2)$$

rewrite in the form $\phi(x)^T \phi(z)$

$$= 1 + 2x_1 z_1 + 2x_2 z_2 + (x_1 z_1)^2 + (x_2 z_2)^2 + 2x_1 z_1 x_2 z_2$$

$\phi_0(z)$ $\phi_1(z)$ - - - - $\phi_5(z)$

$$= \underbrace{(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)^T}_{\phi(x)} \underbrace{(\overset{\phi_0(z)}{1}, \overset{\phi_1(z)}{\sqrt{2}z_1}, \sqrt{2}z_2, z_1^2, z_2^2, \overset{\phi_5(z)}{\sqrt{2}z_1 z_2})}_{\phi(z)}$$

$$= \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{z})$$

- And thus the corresponding kernel $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^6$

Examples

- Generalized polynomial kernel

$$k(\mathbf{x}, \mathbf{x}') = (c + \mathbf{x}^T\mathbf{x}')^M$$

- Gaussian kernel/squared exponential kernel: infinite dimensional space!

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}'||^2)$$

$\underline{\phi}(\underline{x}) \in \mathbb{R}^{M = \infty}$

- Radial basis functions:

$$k(\mathbf{x}, \mathbf{x}') = k(||\mathbf{x} - \mathbf{x}'||^2)$$

Construct new kernels from other kernels

Example Bishop 6.2

$k(\underline{x}, \underline{x}') = \exp(-\|\underline{x}-\underline{x}'\|^2/2)$

derived from

$k_0(\underline{x}, \underline{x}') = \underline{x}^T \underline{x}'$

Each of these kernels correspond to some

$k(\underline{x}, \underline{x}') = \phi(\underline{x})^T \phi(\underline{x}')$

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= c k_1(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') + k_b(\mathbf{x}_b, \mathbf{x}_b') \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') k_b(\mathbf{x}_b, \mathbf{x}_b')
\end{aligned}
$$